On the Rate of Convergence of Fully Connected Deep Neural Regression Estimates

Michael Kohler and Sophie Langer

Presenter: Haoyu Wei

Jan 26, 2021

Multilayer Feedforward Neural Networks

The network structure (L, k) depends on a positive integer L called the *number of hidden layers* and a *width vector* $k = (k_1, \ldots, k_L)^\top \in \mathbb{N}^L$.

Define the activation function $\sigma_{\mathbf{v}}(y_1, \ldots, y_r)^\top = \big(\sigma(y_1 - v_1), \ldots, \sigma(y_r - v_r)\big)^\top$. then a multilayer feedforward neural network with network architecture (L, k) is any function of the form

$$f : \mathbb{R}^d \mapsto \mathbb{R}, \qquad \mathbf{x} \mapsto W_L \sigma_{\mathbf{v}_L} W_{L-1} \sigma_{\mathbf{v}_{L-1}} \cdots W_1 \sigma_{\mathbf{v}_1} W_0 \mathbf{x}, \tag{1}$$

where W_l is a $k_{l+1} \times k_l$ weight matrix and $\mathbf{v}_l \in \mathbb{R}^{k_l}$ is a shift vector.

In the paper, we will use the ReLU activation function

$$\sigma(x) = x \vee 0,$$

with the training data (IID sample) $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$.

Multilayer Feedforward Neural Networks (NN)

A close look at (1) is

$$f(\mathbf{x}) = \sum_{i=1}^{k_L} c_{1,i}^{(L)} f_i^{(L)}(\mathbf{x}) + c_{1,0}^{(L)},$$

for some $c_{1,0}^{(L)}, \ldots, c_{1,k_L}^{(L)} \in \mathbb{R}$ and for $f_i^{(L)}$'s recursively defined by

$$f_i^{(s)}(\mathbf{x}) = \sigma\left(\sum_{j=1}^{k_{s-1}} c_{i,j}^{(s-1)} f_j^{(s-1)}(\mathbf{x}) + c_{i,0}^{(s-1)}\right)$$

for some $c_{i,0}^{(s-1)}, \ldots, c_{i,k_{s-1}}^{(s-1)} \in R$, $s \in 2, \ldots, L$, and

$$f_i^{(i)}(\mathbf{x}) = \sigma\left(\sum_{j=1}^{d} c_{i,j}^{(0)} x_j + c_{i,0}^{(0)}\right)$$

for some $c_{i,0}^{(0)}, \ldots, c_{i,d}^{(0)} \in \mathbb{R}$.

Multilayer Feedforward Neural Networks

Define the space of neural networks with L hidden layers and r neurons per layer as

$$\mathcal{F}(L, r) = \{f : f \text{ is of the form (1) with } k_1 = k_2 = \ldots = k_L = r\}.$$

By minimizing the empirical L_2-risk, our estimator is

$$m_n(\cdot) = \underset{f \in \mathcal{F}(L_n, r_n)}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \big(f(X_i) - Y_i\big)^2. \tag{2}$$

Remark

Since we will not impose any restriction on $c_{i,j}^{(s)}$, we call those networks *fully connected*. Since we can actually take $r = \max k_l$ and some weights are chosen as 0.

Definition ((p, C)-smoothness)

Let $p = q + s$ for some $q \in \mathbb{N}$ and $0 < s \leq 1$. A function $m : \mathbb{R}^d \mapsto \mathbb{R}$ is called (p, C)-smoothness if for every $\alpha = (\alpha_1, \ldots, \alpha_d)^\top \in \mathbb{N}^d$ with $\|\alpha\|_1 = q$,

$$\left| \frac{\partial^q m}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(\mathbf{x}) - \frac{\partial^q m}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}(\mathbf{z}) \right| \leq C \|\mathbf{x} - \mathbf{z}\|_2^s,$$

for any $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$.

The optimal minimax rate of convergence rate in pure nonparametric regression for (p, C)-smooth functions is $n^{-2p/(2p+d)}$ (see [1]), which implies the Curse of Dimensionality.

Remedies

i) Index model

$$m(\mathbf{x}) = \sum_{k=1}^{K} g_k(a_k^\top \mathbf{x})$$

with (p, K)-smooth g_k can achieve the univariate rates of convergence up to some some logarithmic factor (see [2]).

ii) Iteration model

$$m(\mathbf{x}) = g\left(\sum_{l_1=1}^{L_1} g_{l_1}\left(\sum_{l_2=1}^{L_2} g_{l_1,l_2}\left(\dots \sum_{l_r=1}^{L_r} g_{l_1,\dots,l_r}\left(\mathbf{x}_{l_1,\dots,l_r}\right)\right)\right)\right)$$

with (p, C)-smooth univariate functions $g, g_{l_1}, \dots, g_{l_1,\dots,l_r}$ can also reach the univariate rates under some penalized least squares estimating method (see [3]).

Remark

Recall the definition of NN, it can be regard as the mixture of remedy i) and ii). Therefore, it can also circumvent the curse of dimensionality.

What the explicit rate of convergence of a NN?

A single hidden layer NN:

i) [4] gives rate $n^{-1/2}$ (up to some logarithmic factor) when function becomes smoother with increasing d.

ii) [5] shows a rate $n^{(-2p/(2p+d+5)+\varepsilon)}$ for (p, C)-smooth functions with certain cosine squasher as activation function.

Preliminaries and Previous

We introduce a model that contains multilayer NNs.

Definition (generalized hierarchical interaction (see [6]))

Let $d \in \mathbb{N}$, $d^* \in \{1, \ldots, d\}$, and $m : \mathbb{R}^d \mapsto \mathbb{R}$.

a) m satisfies a *generalized hierarchical interaction model of order* d^* *and level* 0, if there exist $a_1, \ldots, a_{d^*} \in \mathbb{R}^d$ and $f : \mathbb{R}^{d^*} \to \mathbb{R}$ such that

$$m(\mathbf{x}) = f(a_1^\top \mathbf{x}, \ldots, a_{d^*}^\top \mathbf{x}).$$

b) m satisfies a *generalized hierarchical interaction model of order* d^* *and level* $l+1$, if there exist $K \in \mathbb{N}$, $g_k : \mathbb{R}^{d^*} \to \mathbb{R}$, and $f_{1,k}, \ldots, f_{d^*,k} : \mathbb{R}^d \to \mathbb{R}$ such that $f_{1,k}, \ldots, f_{d^*,k}$ satisfy a generalized hierarchical interaction model of order d^* and level l and

$$m(\mathbf{x}) = \sum_{k=1}^{K} g_k(f_{1,k}, \ldots, f_{d^*,k}).$$

c) We say that the generalized hierarchical interaction model defined above is (p, C)-smooth, if all functions f and g_k occurring in its definition are (p, C)-smooth

Interaction

The simplest interaction model is

$$m(\mathbf{x}) = \sum_{I \subseteq \{1,\dots,d\}, |I| \leq d^*} m_I(\mathbf{x}_I), \qquad \mathbb{R}^{|I|} \ni \mathbf{x}_I \subseteq \mathbf{x}.$$

If all m_I are (p, C)-smooth for some $p \leq 1$, [7] shows the NN with this structure can achieve a rate of convergence of $n^{-2p/(2p+d^*)}$ (up to some logarithmic factor) independent of d.

NN under Generalized Interaction

Under some relatively strict structure restrictions or using squashing activation function,

i) If *the number of hidden layers suitably depends on the level of the generalized interaction model*, NN can achieve the rate of convergence $n^{-2p/(2p+d^*)}$ (see [6]).

ii) If the activation function is some specific squashing function, the above result can also hold for $p > 1$ with (see [8]).

...

iv) The same convergence rate is shown in [9] using ReLU activation function under proper *network sparsity*.

Network Sparsity

Given a integer $s \asymp \min k_l \times \log n$,

$$\sum_{l=0}^{L} (\|W_l\|_0 + |\mathbf{v}_l|_0) \leq s,$$

which is extremely difficult to implement!

Definition (hierarchical composition model)

Let $d \in \mathbb{N}$ and $m : \mathbb{R}^d \to \mathbb{R}$ and let $\mathcal{P}$ be a subset of $(0, \infty) \times \mathbb{N}$,

a) We say that m satisfies a *hierarchical composition model* of level 0 with order and smoothness constraint $\mathcal{P}$, if there exists a $K \in \{1, \ldots, d\}$ such that

$$m(\mathbf{x}) = x_K, \qquad \forall\, \mathbf{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d.$$

b) We say that m satisfies a hierarchical model of level $l+1$ with order and smoothness constraint $\mathcal{P}$, if there exist $(p, K) \in \mathcal{P}$, $C > 0$, $g : \mathbb{R}^K \to \mathbb{R}$, and $f_1, \ldots, f_K : \mathbb{R}^d \to \mathbb{R}$ satisfy a hierarchical composition model of level l with order and smoothness constraint $\mathcal{P}$ and

$$m(\mathbf{x}) = g\big(f_1(\mathbf{x}), \ldots, f_K(\mathbf{x})\big).$$

Remark

Hierarchical composition model is equivalent to hierarchical composition model under NN!

Hierarchical Composition Model

For $l = 1$ and some order and smoothness constraint $\mathcal{P} \subseteq (0, \infty) \times \mathbb{N}$, the space of hierarchical composition models becomes

$$\begin{aligned}\mathcal{H}(1, \mathcal{P}) = \big\{ & h : \mathbb{R}^d \to \mathbb{R} : h(\mathbf{x}) = g\big(x_{\pi(1)}, \ldots, x_{\pi(K)}\big), \\ & g : \mathbb{R}^K \to \mathbb{R} \text{ is } (p, C)\text{-smooth for some } (p, K) \in \mathcal{P} \\ & \text{and } \pi : \{1, \ldots, K\} \to \{1, \ldots, d\} \big\}.\end{aligned}$$

And for $l > 1$, recursively define

$$\begin{aligned}\mathcal{H}(l, \mathcal{P}) = \big\{ & h : \mathbb{R}^d \to \mathbb{R} : h(\mathbf{x}) = g\big(f_1(\mathbf{x}), \ldots, f_K(\mathbf{x})\big), \\ & g : \mathbb{R}^K \to \mathbb{R} \text{ is } (p, C)\text{-smooth for some } (p, K) \in \mathcal{P} \\ & \text{and } f_j \in \mathcal{H}(l - 1, \mathcal{P}) \big\}.\end{aligned}$$

Define the truncation operator T_β with level $\beta > 0$ as

$$T_\beta u = u \mathbb{I}_{\{|u| \le \beta\}} + \beta \operatorname{sgn}(u) \mathbb{I}_{\{|u| > \beta\}},$$

and $T_\beta \mathcal{F} := \{T_\beta f : f \in \mathcal{F}\}$.

Main Result

Theorem

Let $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ *be IID r.v.s with* $m(\mathbf{x}) = \mathrm{E}[Y \mid X = \mathbf{x}]$. *Suppose:*

1) *the support of* X *is bounded;*
2) $\mathrm{E} \exp\{c_1 Y^2\} < \infty$ *for some constant* $c_1 > 0$;
3) $m \in \mathcal{H}(l, \mathcal{P})$ *for some* $l \in \mathbb{N}$ *and* $\mathcal{P} \subseteq [1, \infty) \times \mathbb{N}$ *with each* g *can be of different smoothness* $p_g = q_g + s_g$ *and of different input dimension* K_g *(*$(p_g, K_g) \in \mathcal{P}$*), and* $p_{\max}, K_{\max} < \infty$.
4) *all partial derivatives of order less than or equal to* q_g *of each* g *in 3) are bounded , i.e.* $\|g\|_{C^{q_g}(\mathbb{R}^{K_g})} \leq c_2$[a].
5) *all* g *in 3) are Lipschitz continuous with Lipschitz constant* $C_{Lip} \geq 1$.

Then let $\tilde{m}_n$ *be defined in (2) for some* $L_n, r_n \in \mathbb{N}$, *and define* $m_n = T_{c_3 \log n} \tilde{m}_n$ *for some sufficiently large.*

---

[a] $\|\cdot\|_{C^q(A)} := \max\{\|\partial^{\mathbf{j}} f\|_{\infty, A} : \|\mathbf{j}\| \leq q, \mathbf{j} \in \mathbb{N}^d\}$

Main Result

Theorem (continue)

Now, two choices for the number of hidden layers and of neurons per layer as follows:

a) Choose $c_4, c_5 > 0$ *sufficiently large and set*

$$L_n = \lceil c_4 \log n \rceil, \qquad r_n = \left\lceil c_5 \max_{(p,K)\in\mathcal{P}} n^{\frac{K}{2(2p+K)}} \right\rceil.$$

Then

$$\mathrm{E} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \, P_X(d\mathbf{x}) \lesssim \log^6 n \times \max_{(p,K)\in\mathcal{P}} n^{-\frac{2p}{2p+K}},$$

holds for sufficiently large n.

b) Choose $c_7, c_8 > 0$ *sufficiently large and set*

$$L_n = \left\lceil c_7 \max_{(p,K)\in\mathcal{P}} n^{\frac{K}{2(2p+K)}} \times \log n \right\rceil, \qquad r_n = r = \lceil c_8 \rceil.$$

Then we also have

$$\mathrm{E} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \, P_X(d\mathbf{x}) \lesssim \log^6 n \times \max_{(p,K)\in\mathcal{P}} n^{-\frac{2p}{2p+K}},$$

Proof Sketch

The basic idea to prove the above theorem can be divided into three steps.

A. Prove for any (p, C)-smooth function f, we can find a NN $f_{NN} \in \mathcal{F}(L, r)$ for some proper $L, r \in \mathbb{N}$ making the maximal error on any cube $[-a, a]^d$ is sufficiently small.

B. Prove for any hierarchical composition model m, we can also find a NN $m_{NN} \in \mathcal{F}(L, r)$ for some proper $L, r \in \mathbb{N}$ making the maximal error on any cube $[-a, a]^d$ is sufficiently small.

C. The NN in Step B can be approximated by the estimator (2) with an arbitrarily small extra error.

Step A: Approximation of smooth functions

Theorem (Approximation Theorem I)

Let $d \in \mathbb{N}$, let $f : \mathbb{R}^d \to \mathbb{R}$ be (p, C)-smooth for some $p = q + s$, $q \in \mathbb{N}$ and $s \in (0, 1]$, and $C > 0$. Suppose $a \geq 1$, $M \in N$ independent of a, and

$$M \geq 2, \qquad M^{2p} \geq c_{10} \left(a \vee \|f\|_{C^q[-a,a]^d}\right)^{4(q+1)},$$

must hold for some sufficiently large constant $c_{10} \geq 1$. Then the in some NN class $\mathcal{F}(L, r)$, we can always find a good NN approximate f enough. Specially,Let $L, r \in N$ such that

(a.i) $L \geq 5 + \lceil \log_4(M^{2p}) \rceil (\lceil \log_2(q \vee d + 1) \rceil + 1)$,

(a.ii) $r \geq 2^d \cdot 64 \cdot \binom{d+q}{d} \cdot d^2 \cdot (q + 1) \cdot M^d$;

or

(b.i) $L \geq 5M^d + \left\lceil \log_4 \left(M^{2p+4d(q+1)} e^{4(q+1)(M^d-1)} \right) \right\rceil \cdot \lceil \log_2(q \vee d + 1) \rceil + \lceil \log_4(M^{2p}) \rceil$,

(b.ii) $r \geq 132 \cdot 2^d \cdot \lceil e^d \rceil \cdot \binom{d+q}{d} \cdot (q + 1) \vee d^2$,

hold. There exists a NN $f_{NN} \in \mathcal{F}(L, r)$ with property that

$$\|f_{NN} - f\|_{\infty,[-a,a]^d} \lesssim \left(a \vee \|f\|_{C^q[-a,a]^d}\right)^{4(q+1)} M^{-2p}.$$

Approximation of Smooth Functions

Directly from the approximation theorem I, one can obtain the following corollary.

Corollary

Let $d \in \mathbb{N}$, $f : \mathbb{R}^d \to \mathbb{R}$ be (p, C)-smooth for some $p = q + s$, $q \in \mathbb{N}$, $s \in (0, 1]$, and $C > 0$. Then for any $a \geq 1$ and $\epsilon > 0$, there exists a fully connected NN f_{NN} with $c_{12}\epsilon^{-d/(2p)}$ parameters such that

$$\|f_{NN} - f\|_{\infty,[-a,a]^d} \leq \epsilon.$$

Comparing to [8] and [9] which require the total number of parameters is $\asymp \epsilon^{-d/p}$ in case of an approximation error ϵ, the above corollary gives a quadratic improvement.

The Proof of Approximation Theorem I

Lemma

Let $d \in \mathbb{N}$, $f : \mathbb{R}^d \to \mathbb{R}$ be (p, C)-smooth for some $p = q + s$, $q \in \mathbb{N}$, $s \in (0, 1]$, and $C > 0$. Then for any $\mathbf{x}_0 \in \mathbb{R}^d$, the Taylor polynomial of total degree q defined by

$$T_{f,q,\mathbf{x}_0}(\mathbf{x}) = \sum_{\mathbf{j} \in \mathbb{N}^d : \|\mathbf{j}\| \leq q} (\partial^{\mathbf{j}} f)(\mathbf{x}_0) \frac{(\mathbf{x} - \mathbf{x}_0)^{\mathbf{j}}}{\mathbf{j}!}$$

satisfies

$$\left| f(\mathbf{x}) - T_{f,q,\mathbf{x}_0}(\mathbf{x}) \right| \leq c_{32} \cdot C \cdot |\mathbf{x} - \mathbf{x}_0|^p$$

for any $\mathbf{x} \in \mathbb{R}^d$ with the constant $c_{32} = c_{32}(q, d)$ depending only on q and d.

Based on this lemma, the coarse idea to prove the above theorem is

1. Divide $[-a, a]^d$ into small enough polytopes and construct proper Taylor polynomial on each polytopes.
2. Recursively define this piecewise Taylor polynomial and then it can be approximated by a NN.

Condition a) and b) may require different approaching details.

The Proof of Approximation Theorem I

Partition $[-a,a)^{d1}$ into M^d and then M^{2d} half-open equivolumn cubes of the form

$$[\mathbf{x},\mathbf{y}) = [\mathbf{x}_1,\mathbf{y}_1) \times \cdots \times [\mathbf{x}_d,\mathbf{y}_d)$$

respectively. Let $\mathcal{P}_1 = \{C_{k,1}\}_{k\in\{1,\ldots,M^d\}}$ and $\mathcal{P}_2 = \{C_{j,2}\}_{j=1}^{M^{2d}}$. Then the piecewise Taylor polynomial

$$T_{f,q,\mathcal{P}_2}(\mathbf{x}) = \sum_{j\in\{1,\ldots,M^{2d}\}} T_{f,q,C_{j,2}}(\mathbf{x}) \cdot \mathbb{I}_{C_{j,2}}(\mathbf{x})$$

satisfies

$$\sup_{\mathbf{x}\in[-a,a)^d} |f(\mathbf{x}) - T_{f,q,\mathcal{P}_2}(\mathbf{x})| \lesssim (2ad)^{2p} \frac{C}{M^{2p}}$$

by the previous lemma.

[1] circumvent overlap

The Proof of Approximation Theorem in case a)

To approximate $f(\mathbf{x})$ by NN our proof follows four key steps:

1. Compute $T_{f,q,\mathcal{P}_2}(\mathbf{x})$ by using recursively defined functions;
2. Approximate the recursive functions by NN on the some inner points in C_{ki} of $\mathcal{P}_2$;
3. Dealing with the boundary lines on each polytope: $f(\mathbf{x})$ times a linear tensorproduct B-spline $w_{\mathcal{P}_2}(\mathbf{x})$.
4. Using $w_{\mathcal{P}_2}(\mathbf{x}) \cdot f(\mathbf{x})$ to approximate $f_{NN}(\mathbf{x})$.

Setting

We order the cubes in $\mathcal{P}_2 = \{C_{ki}\}_{k,i \in \{1,\dots,M^d\}}$ as

$$(C_{ki})_{left} = (C_{i,1})_{left} + \mathbf{v}_k.$$

Step 1: A Recursive Definition of $T_{f,q,\mathcal{P}_2}(\mathbf{x})$

☐ Fining: for $j = 1, \ldots, M^d$, $\mathbf{l} \in \mathbb{N}^d$, and $\|\mathbf{l}\|_1 \leq 1$,

$$\phi_{1,1} = \mathbf{x}, \qquad \phi_{2,1} = \sum_{i=1}^{M^d} \left(C_{i,1}\right)_{left} \cdot \mathbb{I}_{C_{i,1}}(\mathbf{x}), \qquad \phi_{3,1}^{(\mathbf{l},j)} = \sum_{i=1}^{M^d} (\partial^{\mathbf{l}} f)\Big(\left(C_{ji}\right)_{left}\Big) \cdot \mathbb{I}_{C_{i,1}}(\mathbf{x}).$$

Then the group $\{C_{ji}\}_{i=1}^{M^d}$ indexed by j can be defined by

$$\mathcal{A}^{(j)} := \Big\{ \mathbf{x} \in \mathbb{R}^d : -x_k + \left(\phi_{2,1}\right)_k + (\mathbf{v}_j)_k \leq 0,$$
$$x_k - \left(\phi_{2,1}\right)_k - (\mathbf{v}_j)_k - \frac{2a}{M^2}, \quad \forall\, k = 1, \ldots d \Big\}$$

with $\left(C_{ki}\right)_{left} = \left(C_{i,1}\right)_{left} + \mathbf{v}_k$.

☐ Shifting: $\phi_{1,2} = \phi_{1,1}$, and

$$\phi_{2,2} = \sum_{j=1}^{M^d} \left(\phi_{2,1} + \mathbf{v}_j\right) \cdot \mathbb{I}_{\mathcal{A}^{(j)}}\left(\phi_{1,1}\right), \qquad \phi_{3,2}^{(\mathbf{l})} = \sum_{j=1}^{M^d} \phi_{3,1}^{(\mathbf{l},j)} \cdot \mathbb{I}_{\mathcal{A}^{(j)}}\left(\phi_{1,1}\right).$$

Then our Taylor polynomial is define by

$$\phi_{1,3} = \sum_{\mathbf{j} \in \mathbb{N}^d : \|\mathbf{j}\|_1 \leq q} \frac{\phi_{3,2}^{(\mathbf{j})}}{\mathbf{j}!} \cdot \left(\phi_{1,2} - \phi_{2,2}\right)^{\mathbf{j}} = T_{f,q,\mathcal{P}_2}.$$

Melting

For two networks $f \in \mathcal{F}(L_f, r_f)$ and $g \in \mathcal{F}(L_g, r_g)$, then the composed NN $f \circ g$ is contained in $\mathcal{F}(L_f + L_g, r_f \vee r_g)$.

Example

Let

$$f(\mathbf{x}) = \beta_f \sigma(\alpha_f \mathbf{x}), \qquad g(\mathbf{x}) = \beta_g \sigma(\alpha_g \mathbf{x}),$$

then we have

$$f \circ g = f\big(g(\mathbf{x})\big) = \beta_f \sigma\big(\alpha_f \beta_g \sigma(\alpha_g \mathbf{x})\big).$$

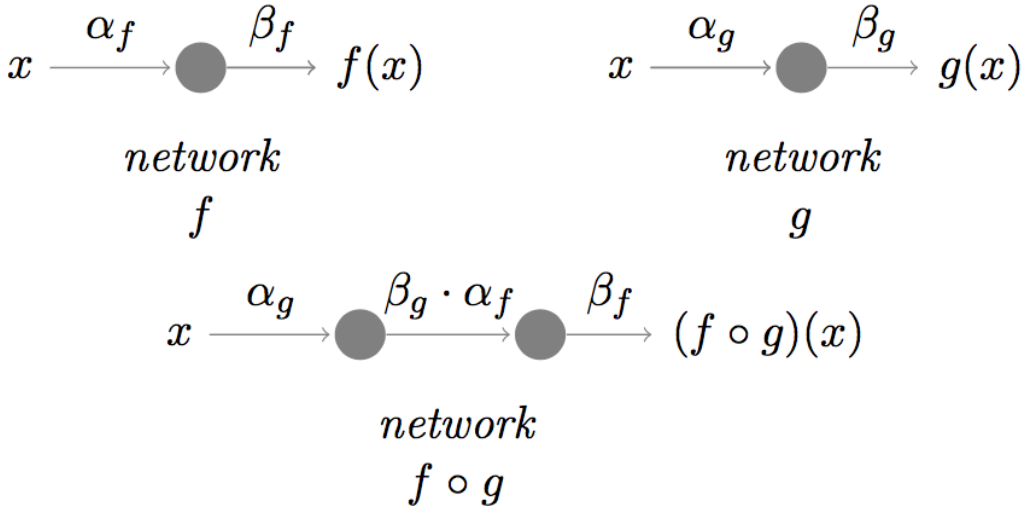


Illustration of the composed network $f \circ g$.

Step 2: Approximating $\phi_{1,3}$ by NN

We will show the following lemma.

Lemma (Approximation Lemma I)

Let $\sigma : \mathbb{R} \to \mathbb{R}$ is the ReLU activation function $\sigma(x) = x \vee 0$. Let $p = q + s$ for some $q \in \mathbb{N}$ and $s \in (0, 1]$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a (p, C)-smooth function. Suppose $a \geq 1$, then there exists $M \in \mathbb{N}$ independent with a sufficiently large, and

$$M^{2p} \gtrsim \left((2a) \vee \|f\|_{C^q[-a,a]^d}\right)^{4(q+1)} \vee C(2ad)^{2p},$$

and a NN $f_{NN,\mathcal{P}_2} \in \mathcal{F}(L, r)$ with

(i) $L = 4 + \lceil \log_2(M^{2p}) \rceil \cdot \lceil \log_2 \left((q+1) \vee 2\right) \rceil$,

(ii) $r = \left[\left(\binom{d+q}{d} + d\right) M^d 2(2 + 2d) + 2d\right] \vee \left[18(q+1)\binom{d+q}{d}\right]$,

such that

$$|f_{NN,\mathcal{P}_2}(\mathbf{x}) - f(\mathbf{x})| \lesssim \left(2a \vee \|f\|_{C^q[-a,a]^d}\right)^{4(q+1)} \frac{1}{M^{2p}}$$

holds for all $\mathbf{x} \in \bigcup_{j=1}^{M^{2d}} \left(C_{j,2}\right)_{1/M^{2p+2}}^{0}$. Besides, the NN value is bounded by

$$|f_{NN,\mathcal{P}_2}(\mathbf{x})| \leq 2e^{2ad} \left(1 \vee \|f\|_{C^q[-a,a]^d}\right)$$

for all $\mathbf{x} \in [-a, a)^d$.

The proof of the Lemma

The following sub-Lemmas show a NN can return approximate x^2 hence xy.

Lemma (sub-lemma 1)

For any $R \in N$ and $a \geq 1$, there exists a NN with ReLU activation function $f_{sq}(x) \in \mathcal{F}(R, 9)$ such that

$$|f_{sq}(x) - x^2| \leq a^2 \cdot 4^{-R}$$

holds for all $x \in [-a, a]$.

Lemma (sub-lemma 2)

For any $R \in N$ and $a \geq 1$, there exists a NN with ReLU activation function $f_{mult}(x, y) \in \mathcal{F}(R, 18)$ such that

$$|f_{mult}(x, y) - xy| \leq 2a^2 \cdot 4^{-R}$$

holds for all $x, y \in [-a, a]$.

Remark

$xy = \frac{1}{4}\left((x+y)^2 - (x-y)^2\right)$.

The proof of sub-lemma 1

1. The tooth function $g : [0,1] \to [0,1]$

$$g(x) = 2x\mathbb{I}_{\{0 \le x \le 1/2\}} + 2(1-x)\mathbb{I}_{\{1/2 < x \le 1\}}$$

satisfies the iterated function

$$g_s(x) = \underbrace{g \circ g \circ \cdots \circ g}_{s}(x) = \begin{cases} 2^s \left(x - \frac{2k}{2^s}\right), & x \in \left[\frac{2k}{2^s}, \frac{2k+1}{2^s}\right], \ k = 0, 1, \ldots, 2^{s-1} - 1, \\ 2^s \left(\frac{2k}{2^s} - 1\right), & x \in \left[\frac{2k-1}{2^s}, \frac{2k}{2^s}\right], \ k = 1, \ldots, 2^{s-1}. \end{cases}$$

2. Then we can show that the linear combination of functions g_s satisfies

$$\left| x - \sum_{s=1}^{R} \frac{g_s(x)}{2^{2s}} - x^2 \right| \le 2^{-2R-2}, \qquad \forall\, x \in [0,1].$$

The proof of sub-lemma 1

3. We can using the following NN to computes $S_R(x) = x - \sum_{s=1}^{R} \frac{g_s(x)}{2^{2s}}$,

$$f_g(x) = 2\sigma(x) - 4\sigma(x - 1/2) + 2\sigma(x - 1), \quad f_{g_s}(x) = \underbrace{f_g(f_g(\ldots(f_g}_{s}(x))) \in \mathcal{F}(s, 3)$$

and

$$\begin{aligned} f_{sq[0,1]} =& f_{id}^R(x) - \frac{1}{2^{2R}} f_{g_R}(x) - f_{id}\left(\frac{1}{2^{2(R-1)}} f_{g_{R-1}}(x)\right. \\ & \left. - f_{id}\left(\frac{1}{2^{2(R-2)}} f_{g_{R-2}}(x) - \cdots - f_{id}\left(\frac{1}{2^2} f_{g_1}(x)\right)\right)\right) = S_R(x) \end{aligned}$$

with $f_{id}(x) = \sigma(x) - \sigma(-x) = x$.

4. For $x \in [-a, a]$, transferring $f_{tran}(z) = \frac{z}{2a} + \frac{1}{2}$ guarantees for any $x \in [-a, a]$,

$$f_{sq}(x) = 4a^2 f_{sq[0,1]}\left(f_{tran}(x)\right) - 2a f_{id}^R(x) - a^2$$

giving $|f_{sq}(x) - x^2| \leq a^2 4^{-R}$.

The proof of the Lemma

We can march a further step. Let $\mathcal{P}_N$ be the linear span of all monomials of the form $\prod_{k=1}^{d} x_k^{r_k}$ with $r_1, \ldots, r_d \in \mathbb{N}$ and $r_1 + \cdots + r_d \leq N$.

Lemma (sub-lemma 3)

Let $m_1, \ldots, m_{\binom{d+N}{d}}$ denote all monomials in $\mathcal{P}_N$ for any $N \in \mathbb{N}_+$. Let $r_1, \ldots, r_{\binom{d+N}{d}} \in \mathbb{R}$, define

$$p\left(\mathbf{x}, y_1, \ldots, y_{\binom{d+N}{d}}\right) = \sum_{i=1}^{\binom{d+N}{d}} r_i y_i \cdot m_i(\mathbf{x}), \quad \mathbf{x} \in [-a, a]^d, \quad y_i \in [-a, a],$$

and set $\bar{r}(p) = \max |r_i|$. Then for any $a \geq 1$, and $R \geq \log_4\left(2 \cdot 4^{2(N+1)} a^{2(N+1)}\right)$, a NN $f_p\left(\mathbf{x}, y_1, \ldots, y_{\binom{d+N}{d}}\right) \in \mathcal{F}(L, r)$ with ReLU activation function, and $L = R\lceil \log_2(N+1) \rceil$, $r = 18(N+1)\binom{d+N}{d}$ exists, such that

$$\left| f_p\left(\mathbf{x}, y_1, \ldots, y_{\binom{d+N}{d}}\right) - p\left(\mathbf{x}, y_1, \ldots, y_{\binom{d+N}{d}}\right) \right| \lesssim \bar{r}(p) a^{4(N+1)} \cdot 4^{-R},$$

for all $x_i, y_j \in [-a, a]$, and the constant above only depends on d and N.

The proof of sub-lemma 3

1. From sub-lemma 2, there exists a NN with ReLU activation function $f_m(y, \mathbf{x}) = \mathcal{F}\big(R\lceil \log_2(N+1), 18(N+1)\rceil\big)$ achieves an approximation error

$$|f_m(\mathbf{x}, y) - ym(\mathbf{x})| \leq 4 \cdot 4^{4(N+1)} \cdot a^{4(N+1)} \cdot (N+1) \cdot 4^{-R}.$$

2. Conclude that

$$\left| p\left(\mathbf{x}, y_1, \ldots, y_{\binom{d+N}{d}}\right) - \sum_{i=1}^{\binom{d+N}{d}} r_i f_{m_i}(\mathbf{x}, y_i) \right| \leq \binom{d+N}{d} \bar{r}(p) 4 \cdot 4^{4(N+1)} a^{4(N+1)} (N+1) \cdot 4^{-R}.$$

Remark

From sub-lemma 3, it can be shown that a NN $f_p\left(\mathbf{z}, y_1, \ldots, y_{\binom{d+q}{q}}\right) \in \mathcal{F}\left(B_{M,p}\lceil \log_2((q+1) \vee 2)\rceil, 18(q+1)\binom{d+q}{q}\right)$ satisfying

$$\left| f_p\left(\mathbf{z}, y_1, \ldots, y_{\binom{d+q}{q}}\right) - p\left(\mathbf{z}, y_1, \ldots, y_{\binom{d+q}{q}}\right) \right| \lesssim \bar{r}(p) \left(2a \vee \|f\|_{C^q[-a,a]^d}\right)^{4(q+1)} \cdot 4^{-B_{M,p}}$$

for all z_i, y_j contained in $\left[-2a \vee \|f\|_{C^q[-a,a]^d}, 2a \vee \|f\|_{C^q[-a,a]^d}\right]$ where $\log_4\left(2 \cdot 4^{2(q+1)} 2a \vee \|f\|_{C^q[-a,a]^d}\right) \leq B_{M,p} \in \mathbb{N}$.

The proof of the Lemma

The advantage of ReLU activating that it is zero in case of negative input can approximates the multidimensional indicator function (multiplied by a additional factor).

Lemma (sub-lemma 4)

Let $R \in \mathbb{N}$, $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ *with* $b_i - a_i \geq 2/R$, *and let*

$$K_{1/R} = \{\mathbf{x} \in \mathbb{R}^d : x_i \notin [a_i, a_i + 1/R) \cup (b_i - 1/R, b_i),\ i = 1, \ldots, d\}.$$

a) Then the network

$$f_{ind,[\mathbf{a},\mathbf{b})}(\mathbf{x}) = \sigma\left(1 - R \sum_{i=1}^{d} \Big(\sigma(a_i + 1/R - x_i) + \sigma(x_i - b_i + 1/R)\Big)\right)$$

of class $\mathcal{F}(2, 2d)$ *satisfies* $f_{ind,[\mathbf{a},\mathbf{b})}(\mathbf{x}) = \mathbb{I}_{[\mathbf{a},\mathbf{b})}(\mathbf{x})$ *for* $\mathbf{x} \in K_{1/R}$ *and*

$$\left|f_{ind,[\mathbf{a},\mathbf{b})}(\mathbf{x}) - \mathbb{I}_{[\mathbf{a},\mathbf{b})}(\mathbf{x})\right| \leq 1,$$

for all $\mathbf{x} \in \mathbb{R}^d$.

The proof of the Lemma

Lemma (sub-Lemma 4, continue)

b) Let $|s| \leq R$. Then the NN

$$f_{test}(\mathbf{x}, \mathbf{a}, \mathbf{b}, s) = \sigma\Big(f_{id}(s) - R^2 \sum_{i=1}^{d} \Big(\sigma(a_i + 1/R - x_i) + \sigma(x_i - b_i + 1/R) \Big) \Big)$$
$$- \sigma\Big(- f_{id}(s) - R^2 \sum_{i=1}^{d} \Big(\sigma(a_i + 1/R - x_i) + \sigma(x_i - b_i + 1/R) \Big) \Big)$$

of the class $\mathcal{F}(2, 2(2d+2))$ satisfies $f_{test}(\mathbf{x}, \mathbf{a}, \mathbf{b}, s) = s\mathbb{I}_{[\mathbf{a},\mathbf{b})}$ for $\mathbf{x} \in K_{1/R}$ and

$$\left| f_{test}(\mathbf{x}, \mathbf{a}, \mathbf{b}, s) - s\mathbb{I}_{[\mathbf{a},\mathbf{b})} \right| \leq |s|,$$

for any $\mathbf{x} \in \mathbb{R}^d$.

The proof of this lemma is direct.

The Proof of the Lemma

With assistance of sub-lemma 1-4, we can prove the Approximation Lemma I (the Lemma in Step 2).

I. How recursively defined function $\phi_{1,3}$ can be approximated by NNs?

Construct the NN by following

$$\widehat{\phi}_{1,1} = f_{id}^2, \quad \widehat{\phi}_{2,1} = \sum_{i=1}^{M^d} \left(C_{i,1}\right)_{left} \cdot f_{ind,C_{i,1}}, \quad \widehat{\phi}_{3,1}^{(\mathbf{l},j)} = \sum_{i=1}^{M^d} (\partial^{\mathbf{l}} f)\left((C_{ji})_{left}\right) \cdot f_{ind,C_{i,1}}.$$

$$\widehat{\phi}_{1,2} = f_{id}^2 \circ \widehat{\phi}_{1,1}, \quad \left(\widehat{\phi}_{2,2}\right)_i = \sum_{j=1}^{M^d} f_{test}\left(\hat{\phi}_{1,1}, \hat{\phi}_{2,1} + \mathbf{v}_j, \hat{\phi}_{2,1} + \mathbf{v}_j + 2a/M^2 \cdot \mathbf{1}, \widehat{\phi}_{2,1}^{(i)} + (\mathbf{v}_j)_i\right),$$

$$\widehat{\phi}_{3,2}^{(\mathbf{l})} = \sum_{j=1}^{M^d} f_{test}\left(\hat{\phi}_{1,1}, \hat{\phi}_{2,1} + \mathbf{v}_j, \hat{\phi}_{2,1} + \mathbf{v}_j + 2a/M^2 \cdot \mathbf{1}, \widehat{\phi}_{3,1}^{(\mathbf{l},j)}\right).$$

Then we choose $\mathbf{l}_1, \ldots, \mathbf{l}_{\binom{d+q}{d}}$ such that

$$\left\{\mathbf{l}_1, \ldots, \mathbf{l}_{\binom{d+q}{d}}\right\} = \left\{(s_1, \ldots, s_d) \in \mathbb{N}^d : s_1 + \cdots + s_d \leq q\right\}.$$

The Proof of the Lemma

I. How recursively defined function $\phi_{1,3}$ can be approximated by NNs? (continue)

Then the value of $\phi_{1,3}$ can be computed by

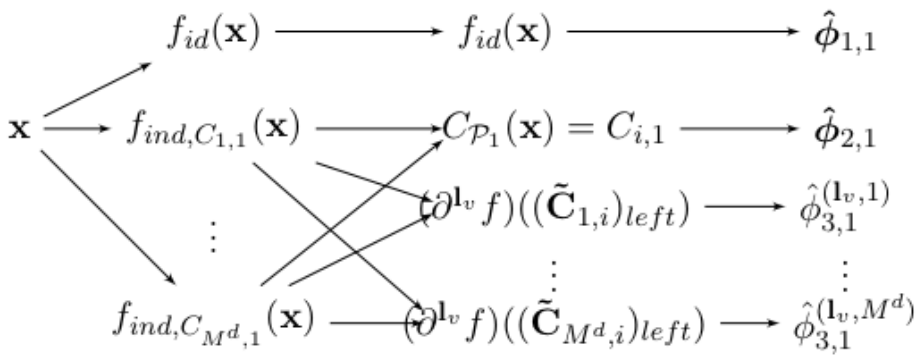
$$\widehat{\phi}_{1,3} = f_p\left(\widehat{\phi}_{1,2} - \widehat{\phi}_{2,2}, \widehat{\phi}_{3,2}^{(\mathbf{l}_1)}, \ldots, \widehat{\phi}_{3,2}^{(\mathbf{l}_{\binom{d+q}{d}})}\right).$$

Hence, we construct a NN to approximate $\phi_{1,3}$.

For computing the number of hidden layer and neurons, we need to note that

$$\left(\widehat{\phi}_{1,1}, \widehat{\phi}_{2,1}, \widehat{\phi}_{3,1}^{(\mathbf{l}_v,1)}, \ldots, \widehat{\phi}_{3,1}^{(\mathbf{l}_v,M^d)}\right)$$

is constructed by the following structure



Thus, it needs $L_1 = 2$ hidden layers and $r_1 = 2d + d \cdot M^d \cdot 2d + M^d\binom{d+q}{d} \cdot 2d$ neurons per layer.

The Proof of the Lemma

I. How recursively defined function $\phi_{1,3}$ can be approximated by NNs? (continue)

Furthermore, we can conclude that $(\widehat{\phi}_{1,2}, \widehat{\phi}_{2,2}, \widehat{\phi}_{3,2}^{(\mathbf{l}_v)})$ needs $L_2 = L_1 + 2 = 4$ hidden layers and

$$r_2 = r_1 \vee \left(2d + d \cdot M^d \cdot 2 \cdot (2d+2) + \binom{d+q}{d} \cdot M^d \cdot 2 \cdot (2d+2) \right)$$
$$= 2d + \left(d + \binom{d+q}{d} \right) \cdot M^d 2(2d+2)$$

neurons per layers. Finally, we have that $\widehat{\phi}_{1,3}$ lies in $\mathcal{F}\big(4 + B_{M,p}\lceil \log_2((q+1) \vee 2) \rceil, r\big)$ with

$$r = r_2 \vee 18(q+1)\binom{d+q}{d}.$$

We set $f_{NN,\mathcal{P}_2} = \widehat{\phi}_{1,3}$.

The Proof of the Lemma

II. What the error of the NN $f_{NN,\mathcal{P}_2}(\mathbf{x})$ far away from the boundaries?

We consider the case that

$$B_M \geq M^{2p+2}, \qquad \mathbf{x} \in \bigcup_{k=1}^{M^{2d}} \left(C_{k,2}\right)^0_{1/M^{2p+2}}.$$

From sub-lemma 4, we know that the NN $\widehat{\phi}_{1,1}, \widehat{\phi}_{2,1}, \widehat{\phi}_{3,1}^{(\mathbf{l}_v,1)}, \ldots, \widehat{\phi}_{3,1}^{(\mathbf{l}_v,M^d)}$ and the NN $\widehat{\phi}_{1,2}, \widehat{\phi}_{2,2}, \widehat{\phi}_{3,2}^{(\mathbf{l}_v)}$ compute the corresponding functions without a error. It follows that

$$\left|\widehat{\phi}_{1,2} - \widehat{\phi}_{2,2}\right| = |\mathbf{x} - \phi_{2,2}| \leq 2a, \qquad \left|\widehat{\phi}_{3,2}\right| = |\phi_{3,2}| \leq \|f\|_{C^q[-a,a]^d}.$$

By choosing $B_{M,p} = \lceil \log_4(M^{2p}) \rceil$,

$$|f_{NN,\mathcal{P}_2} - T_{f,q,\mathcal{P}_2}| = \left|\widehat{\phi}_{1,3} - \phi_{1,3}\right| \lesssim \left(2a \vee \|f\|_{C^q[-a,a]^d}\right)^{4(q+1)} \frac{1}{M^{2p}}.$$

And the value of the NN is bounded by

$$|f_{NN,\mathcal{P}_2}(\mathbf{x})| \leq 2\Big(1 \vee \sup_{[-a,a]^d} |f(\mathbf{x})|\Big).$$

The Proof of the Lemma

II. What the upper bound of the NN $f_{NN,\mathcal{P}_2}(\mathbf{x})$ close boundaries?

For the case that

$$\mathbf{x} \in \bigcup_{k=1}^{M^{2d}} C_{k,2}/(C_{k,2})^0_{1/M^{2p+2}}.$$

By the construction, for $\mathbf{x} \in C_{i,1}$, sub-lemma 4 guarantees

$$\left|\widehat{\phi}_{3,1}^{(\mathbf{l},j)}\right| \leq \left|(\partial^{\mathbf{l}} f)(C_{ji})_{left}\right|, \qquad \left|(\widehat{\phi}_{2,1})_s\right| \leq a, \qquad j = 1, \ldots, M^d, \ s = 1, \ldots, d.$$

which implies

$$\left|\widehat{\phi}_{3,2}^{(\mathbf{l})}\right| \leq \|f\|_{C^q[-a,a]^d}, \qquad \left|(\widehat{\phi}_{2,2})_s\right| \leq a.$$

Hence

$$|f_{NN,\mathcal{P}_2}| \leq |f_p(\cdot) - p(\cdot)| + |p(\cdot)| \leq 1 + e^{2ad}\|f\|_{C^q[-a,a]^d}.$$

Remark

Note that in Approximation Lemma I we do not control the error when $\mathbf{x}$ is close to the boundaries. This will be remained to Step 3.

Step 3: Tackling Boundaries

We will use construct a NN to approximate $w_{\mathcal{P}_2}(\mathbf{x}) \cdot f(\mathbf{x})$, where

$$w_{\mathcal{P}_2}(\mathbf{x}) = \prod_{j=1}^{d} \left(1 - \frac{M}{a^2} \left| (C_{\mathcal{P}_2}(\mathbf{x}))_{left,j} + \frac{a}{M^2} - x_j \right| \right)_+$$

is a linear tensorproduct B-spline which takes its maximum value at the center of $C_{\mathcal{P}_2}(\mathbf{x})$ and which vanishes outside of $C_{\mathcal{P}_2}(\mathbf{x})$.

Then $w_{\mathcal{P}_2}(\mathbf{x})$ is close to zero when $\mathbf{x}$ closing to the boundary of $C_{\mathcal{P}_2}(\mathbf{x})$. In fact, we have

$$w_{\mathcal{P}_2}(\mathbf{x}) \leq \frac{1}{M^{2p}}, \qquad \mathbf{x} \in \bigcup_{k=1}^{M^{2d}} C_{k,2} / (C_{k,2})^0_{1/M^{2p+2}}.$$

Step 3: Tackling Boundaries Lines

Lemma (Approximation Lemma II)

Assume the conditions are the same as that in Approximation Lemma I. Define $w_{\mathcal{P}_2}$ as previous. Then there exists a NN $f_{NN} \in \mathcal{F}(L, r)$ with

$$L = 5 + \lceil \log_4(M^{2p}) \rceil \cdot \lceil \log_2((q+1) \vee 2) \rceil, \qquad r = 64 \binom{d+q}{d} d^2 (q+1) M^d,$$

such that

$$|f_{NN}(\mathbf{x}) - w_{\mathcal{P}_2}(\mathbf{x}) f(\mathbf{x})| \lesssim \left(2a \vee \|f\|_{C^q[-a,a]^d}\right)^{4(q+1)} \frac{1}{M^{2p}}$$

holds for any $\mathbf{x} \in [-a, a)^d$.

The proof of this lemma is essentially finding a NN to approximate $w_{\mathcal{P}_2}(\mathbf{x})$, and then a NN to approximate the product.

Some basic sub-lemmas for proving Approximate Lemma II

Lemma (sub-lemma 5)

For any $R \in \mathbb{N}$ and $a \geq 1$, there exists a NN $f_{mult,d} \in \mathcal{F}(R\lceil \log_2 d \rceil, 18d)$ with ReLU activation function such that

$$\left| f_{mult,d}(\mathbf{x}) - \prod_{i=1}^{d} x_i \right| \leq 4^{4d+1} \cdot a^{4d} \cdot d \cdot 4^{-R}$$

holds for any $\mathbf{x} \in [-a, a]^d$.

Obviously, sub-lemma 5 is actually a extension of sub-lemma 2, and its proof is based on using sub-lemma 2 recursively.

Some basic sub-lemmas for proving Approximate Lemma II

Lemma (sub-lemma 6)

Let $a \geq 1$ and $M \geq d4^{2d+1}$, and $\mathcal{P}_2$ and $w_{\mathcal{P}_2}$ defined as previous. Then there exist a NN with ReLU activation function,

$$f_{w_{\mathcal{P}_2}}(\mathbf{x}) \in \mathcal{F}\left(5 + \lceil \log_4(M^{2p}) \rceil \lceil \log_2 d \rceil, r\right),$$

where

$$r = (18d) \vee (2d + d \cdot M^d \cdot 2 \cdot (2 + 2d))$$

such that

$$\left| f_{w_{\mathcal{P}_2}}(\mathbf{x}) - w_{\mathcal{P}_2}(\mathbf{x}) \right| \leq 4^{4d+1} \cdot d \cdot \frac{1}{M^{2p}}$$

for any $\mathbf{x} \in \bigcup_{i=1}^{M^{2d}} (C_{i,2})^0_{1/M^{2p+2}}$ and

$$|f_{w_{\mathcal{P}_2}}(\mathbf{x})| \leq 2$$

for all $\mathbf{x} \in [-a, a)^d$.

The Proof of sub-lemma 6

Note that the j-th factor in $w_{\mathcal{P}_2}$ can be decomposed as

$$\left(\frac{M}{a^2}\left(x_j-\left(C_{\mathcal{P}_2}(\mathbf{x})\right)_{left,j}\right)\right)_+-2\left(\frac{M}{a^2}\left(x_j-\left(C_{\mathcal{P}_2}(\mathbf{x})\right)_{left,j}-\frac{a}{M^2}\right)\right)_+$$
$$+\left(\frac{M}{a^2}\left(x_j-\left(C_{\mathcal{P}_2}(\mathbf{x})\right)_{left,j}-\frac{2a}{M^2}\right)\right)$$

with every component can be easily computed by applying the ReLU activation function, and then applying sub-lemma 5 for the product.

I. Compute the value of $(C_{\mathcal{P}_2}(\mathbf{x}))_{left}$ by $\widehat{\phi}_{2,2}$.

II. The the j-factor in $w_{\mathcal{P}_2}$ can be approximated by

$$f_{w_{\mathcal{P}_2},j}(\mathbf{x})=\sigma\left(\frac{M}{a^2}\left(\left(\widehat{\phi}_{1,2}\right)_j-\left(\widehat{\phi}_{2,2}\right)_j\right)\right)-2\sigma\left(\frac{M}{a^2}\left(\left(\widehat{\phi}_{1,2}\right)_j-\left(\widehat{\phi}_{2,2}\right)_j-\frac{a}{M^2}\right)\right)$$
$$+\sigma\left(\frac{M}{a^2}\left(\left(\widehat{\phi}_{1,2}\right)_j-\left(\widehat{\phi}_{2,2}\right)_j-\frac{2a}{M^2}\right)\right).$$

The Proof of sub-lemma 6

III. The product of $w_{\mathcal{P}_2,j}$ can be computed by the NN in sub-lemma 5, i.e.

$$f_{w_{\mathcal{P}_2}}(\mathbf{x}) = f_{mult,d}\left(f_{w_{\mathcal{P}_2},1}(\mathbf{x}), \ldots, f_{w_{\mathcal{P}_2},d}(\mathbf{x})\right),$$

which satisfies all the requirements in sub-lemma 6.

Remark

Both the NN $f_{NN,\mathcal{P}_2}$ in Approximation Lemma I and $f_{w_{\mathcal{P}_2}}$ in sub-lemma 6 can only approximate the original function in case that $\mathbf{x}$ is not close to the boundaries of the grids of $\mathcal{P}_2$. We need an auxiliary NN to control the error in case that

$$\mathbf{x} \in \bigcup_{k=1}^{M^d} C_{k,2}/(C_{k,2})^0_{1/M^{2p+2}}.$$

From some prospective, this auxiliary NN actually checks the position of $\mathbf{x}$, we will call it *check* NN.

Some basic sub-lemmas for proving Approximate Lemma II

Lemma (sub-lemma 7)

Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be the partitions defined as previous and $M \in \mathbb{N}$. Then there exists a NN with ReLU activation function

$$f_{check,\mathcal{P}_2}(\mathbf{x}) \in \mathcal{F}\big(5, 2d + (4d^2 + 4d) \cdot M^d\big)$$

satisfying

$$f_{check,\mathcal{P}_2}(\mathbf{x}) = \mathbb{I}_{\bigcup_{i=1}^{M^{2d}} C_{i,2} / (C_{i,2})^0_{1/M^{2p+2}}}(\mathbf{x})$$

for any $\mathbf{x} \notin \bigcup_{i=1}^{M^{2d}} (C_{i,2})^0_{1/M^{2p+2}} \,/\, (C_{i,2})^0_{2/M^{2p+2}}$ and

$$f_{check,\mathcal{P}_2}(\mathbf{x}) \in [0, 1]$$

for any $\mathbf{x} \in [-a, a]^d$.

Remark

The $f_{check,\mathcal{P}_2}$ actually check whether $\mathbf{x}$ is close to the boundaries.

The proof of sub-lemma 7

I. Construct a NN for checking whether $\mathbf{x}$ is close to the boundaries of $\mathcal{P}_1$?

Our NN can approximates in the first 2 hidden layers the function

$$f_1(\mathbf{x}) = \mathbb{I}_{\bigcup_{i=1}^{M^d} C_{i,1}/(C_{i,1})^0_{1/M^{2p+2}}}(\mathbf{x}) = 1 - \sum_{i=1}^{M^d} \mathbb{I}_{(C_{i,1})^0_{1/M^{2p+2}}}(\mathbf{x})$$

by

$$\widehat{f}_1(\mathbf{x}) = 1 - \sum_{k=1}^{M^d} f_{ind,(C_{k,1})^0_{1/M^{2p+2}}}(\mathbf{x}).$$

The proof of sub-lemma 7

II. Construct a NN for checking whether $\mathbf{x}$ is close to the boundaries of $\mathcal{P}_2$?

a) Compute the position of $(C_{\mathcal{P}_2}(\mathbf{x}))_{left}$ by using $\widehat{\phi}_{2,1} = \sum_{i=1}^{M^d} \left(C_{i,1}\right)_{left} \cdot f_{ind,C_{i,1}}$.

b) Describe the cubes $(C_{ji})_{1/M^{2p+2}}^0$ index by j by[2]

$$(\mathcal{A}^{(j)})_{1/M^{2p+2}}^0 = \Big\{ \mathbf{x} \in \mathbb{R}^d : -x_k + (\phi_{2,1})_k + (\mathbf{v}_j)_k + \frac{1}{M^{2p+2}} \leq 0$$
$$x_k - (\phi_{2,1})_k - (\mathbf{v}_j)_k - \frac{2a}{M^2} + \frac{1}{M^{2p+2}} < 0, \ \forall k = 1, \ldots, d \Big\}.$$

Then by sub-lemma 4 b) the function

$$f_2(\mathbf{x}) = \mathbb{I}_{\bigcup_{j=1}^{M^d} C_{ji}/(C_{ji})_{1/M^{2p+2}}^0}(\mathbf{x}) = 1 - \sum_{j=1}^{M^d} \mathbb{I}_{(C_{ji})_{1/M^{2p+2}}^0}(\mathbf{x})$$

can be approximated by

$$\widehat{f}_2(\mathbf{x}) = 1 - \sum_{j=1}^{M^d} f_{test}\left(f_{id}(\mathbf{x})^2, \widehat{\phi}_{2,1} + \mathbf{v}_j + \frac{1}{M^{2p+2}}\mathbf{1}, \widehat{\phi}_{2,1} + \mathbf{v}_j \left(\frac{2a}{M^2} - \frac{1}{M^{2p+2}} \right) \mathbf{1}, 1 \right),$$

[2] the same technique that we used in Step 1!

The proof of sub-lemma 6

II. Construct a NN for checking whether $\mathbf{x}$ is close to the boundaries of $\mathcal{P}_2$? (continue)

c) Combining f_1 and f_2. The NN

$$f_{check,\mathcal{P}_2}(\mathbf{x}) = 1 - \sigma\Big(1 - \widehat{f}_2(\mathbf{x}) - f_{id}^2\big(\widehat{f}_1(\mathbf{x})\big)\Big) \in [0,1]$$

is contained in $\mathcal{F}(5,r)$ with

$$r = \big(2d + d \cdot M^d \cdot 2d + M^d \cdot 2d\big) \vee \big(M^d 2(2+2d)+2\big) \leq 2d + (4d^2+4d)M^d.$$

Then we can verify $f_{check,\mathcal{P}_2}$ is actually what we want to find in the sub-lemma 7.

The Proof of Approximation Lemma II

The sub-lemma 6 and sub-lemma 7 lead to a NN approximating $w_{\mathcal{P}_2}(\mathbf{x}) \cdot f(\mathbf{x})$ for any $\mathbf{x} \in [-a, a)^d$ by defining

$$f_{mult}\big(f_{w_{\mathcal{P}_2}}(\mathbf{x}), f_{net,\mathcal{P}_2,true}(\mathbf{x})\big) \approx f_{w_{\mathcal{P}_2}}(\mathbf{x}) \cdot f_{net,\mathcal{P}_2,true}(\mathbf{x})$$

with

$$\begin{aligned} f_{net,\mathcal{P}_2,true}(\mathbf{x}) =& \sigma\big(f_{net,\mathcal{P}_2}(\mathbf{x}) - B_{true} \cdot f_{check,\mathcal{P}_2}(\mathbf{x})\big) \\ & - \sigma\big(- f_{net,\mathcal{P}_2}(\mathbf{x}) - B_{true} \cdot f_{check,\mathcal{P}_2}(\mathbf{x})\big). \end{aligned}$$

a) If $\mathbf{x}$ is away from the boundaries, $f_{check,\mathcal{P}_2}$ is 0, thus $f_{net,\mathcal{P}_2,true}(\mathbf{x}) = f_{net,\mathcal{P}_2}(\mathbf{x})$.

b) If $\mathbf{x}$ is close to the boundaries, $f_{check,\mathcal{P}_2}(\mathbf{x})$ is 1, thus $f_{net,\mathcal{P}_2,true}(\mathbf{x})$ is 0.

Remark

The existence of $w_{\mathcal{P}_2}$ guarantees when $\mathbf{x}$ is close to boundaries, the product $w_{\mathcal{P}_2}(\mathbf{x}) \cdot f(\mathbf{x})$ is also close to zero!

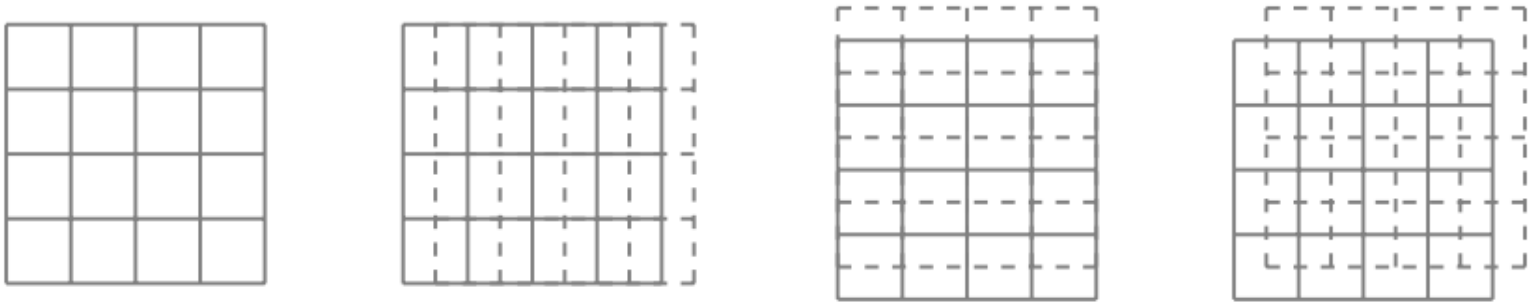
Step 4: Using $w_{\mathcal{P}_2}(\mathbf{x}) \cdot f(\mathbf{x})$ to approximate $f_{NN}(\mathbf{x})$

We will show that

$$\sup_{\mathbf{x}\in[-a/2,a/2]^d} |f(\mathbf{x}) - f_{net,wide}(\mathbf{x})| \lesssim \left((2a) \vee \|f\|_{C^q[-a,a]^d}\right)^{4(q+1)} \frac{1}{M^{2p}}.$$

The half cube allows us to do some small sight for the partition $\mathcal{P}_1$ and $\mathcal{P}_2$.

If one of the components it shifted by a/M^2, there exist 2^d shifts. We set $\mathcal{P}_{1,1} = \mathcal{P}_1$ and $\mathcal{P}_{2,1} = \mathcal{P}_1$, and define each $v = \{2, \ldots, 2^d\}$ partitions $\mathcal{P}_{1,v}$ and $\mathcal{P}_{2,v}$.



Example: 2^2 different partitions in case $d = 2$.

Step 4: Using $w_{\mathcal{P}_2}(\mathbf{x}) \cdot f(\mathbf{x})$ to approximate $f_{NN}(\mathbf{x})$

Under v-th partition, set the NN and the B-spline in Approximation Lemma II as $f_{NN,v}$ and w_v respectively. Then for the wide NN $f_{NN,wide}(\mathbf{x}) = \sum_{v=1}^{2^d} f_{NN,v}(\mathbf{x}) \in \mathcal{F}(L, r)$,

$$
\begin{aligned}
|f_{NN,wide}(\mathbf{x}) - f(\mathbf{x})| &= \left| \sum_{v=1}^{2^d} f_{NN,v}(\mathbf{x}) - \sum_{v=1}^{2^d} w_v(\mathbf{x}) \cdot f(\mathbf{x}) \right| \\
&\leq \sum_{v=1}^{2^d} |f_{NN,v}(\mathbf{x}) - w_{\mathcal{P}_2,v}(\mathbf{x}) f(\mathbf{x})| \\
&\lesssim \left(2a \vee \|f\|_{C^q[-a,a]^d}\right)^{4(q+1)} \frac{1}{M^{2p}}.
\end{aligned}
$$

And the explicit L and r can be computed by the construction.

The Proof of Approximation Theorem in case b)

The basic idea to prove case b) is the same as that in case a), while the construction should be a slightly different, since we should use more iteration and less neurons for building a *deep* rather than a *wide* NN!

Setting

In case b), we order $\mathcal{P}_2 = \{C_{ki}\}_{k,i \in \{1,\dots,M^d\}}$ in another orders: $(C_{1i})_{left} = (C_{i,1})_{left}$, and

$$(C_{ki})_{left} = (C_{(k-1)\,i})_{left} + \mathbf{v}_k.$$

Remark

Since using the original indicator form of $T_{f,q,\mathcal{P}_2}$ requires many neurons, we need another construction to compute $T_{f,q,\mathcal{P}_2}$, by artificially giving a $\widehat{T}_{f,q,\mathcal{P}_2}$ enough close to $T_{f,q,\mathcal{P}_2}$ which requires many layers to approximate.

The proof of case b) is also divided into four steps. We only illustrate the Step 1* - 2* as Step 3* - 4* are basically the same as Step 3 - Step 4.

Step 1*-1: How to construct $\widehat{T}_{f,q,\mathcal{P}_2}(\mathbf{x})$?

I. Compute $(C_{\mathcal{P}_1}(\mathbf{x}))_{left}$ and $(\partial^{\mathbf{l}} f)((C_{i,1})_{left})$ and suitably define numbers

$$b_{ki}^{(\mathbf{l})} \in \mathbb{Z}, \qquad \left|b_{ki}^{(\mathbf{l})}\right| \leq e^d + 1, \quad k = 1, \ldots, M^d.$$

Now fix $\mathbf{x} \in C_{i,1}$.

II. Give NN successively calculate approximation

$$(\partial^{\mathbf{l}} \widehat{f})\big((C_{ki})_{left}\big), \qquad k = 1, \ldots, M^d.$$

of $(\partial^{\mathbf{l}} f)((C_{ki})_{left})$.

III. Give

$$\widehat{T}_{f,q,\mathcal{P}_2}(\mathbf{x}) = \sum_{\mathbf{l} \in \mathbb{N}^d : \|\mathbf{l}\|_1 \leq q} \frac{(\partial^{\mathbf{l}} \widehat{f})((C_{\mathcal{P}_2}(\mathbf{x}))_{left})}{\mathbf{l}!} \big(\mathbf{x} - (C_{\mathcal{P}_2}(\mathbf{x}))_{left}\big)^{\mathbf{l}}.$$

A close look at sub-step II

Start with $(\partial^{\mathbf{l}} \widehat{f})\big((C_{1i})_{left}\big) = (\partial^{\mathbf{l}} \widehat{f})\big((C_{\mathcal{P}_1}(\mathbf{x}))_{left}\big)$, then recursively compute

$$(\partial^{\mathbf{l}} \widehat{f})\big((C_{(k+1)\,i})_{left}\big) = \sum_{\mathbf{j} \in \mathbb{N}^d : \|\mathbf{j}\|_1 \leq q - \|\mathbf{l}\|_1} \frac{(\partial^{\mathbf{l}+\mathbf{j}} \widehat{f})\big((C_{ki})_{left}\big)}{\mathbf{j}!} \big((C_{(k+1)\,i})_{left} - (C_{ki})_{left}\big)$$
$$+ b_{ki}^{(\mathbf{l})} \cdot c_{46} \cdot \left(\frac{2a}{M^2}\right)^{p - \|\mathbf{l}\|_1},$$

where

$$c_{46} = C \cdot d^p \cdot \max\{c_{32}(q,d), \ldots, c_{32}(0,d)\}$$

with c_{32} is the constant in Taylor lemma.

A close look at sub-step I and III

Remark

It is easy to check that the choice of $b_{ki}^{(\mathbf{l})}$ guarantees under this recursive definition, we recursively have

$$\left|(\partial^{\mathbf{l}}\widehat{f})\big((C_{(k+1)\,i})_{left}\big) - (\partial^{\mathbf{l}}f)\big((C_{(k+1)\,i})_{left}\big)\right| \leq c_{46} \cdot \left(\frac{2a}{M^2}\right)^{p-\|\mathbf{l}\|_1}.$$

Besides, for $b_{ki}^{(\mathbf{l})}$, we have

$$b_i^{(\mathbf{l})} = \sum_{k=1}^{M^d-1} \left(b_{ki}^{(\mathbf{l})} + \lceil e^d \rceil + 2\right) \cdot (4 + 2\lceil e^d \rceil)^{-k} \in [0,1].$$

Furthermore, we have

$$\left|\widehat{T}_{f,q,\mathcal{P}_2}(\mathbf{x}) - T_{f,q,\mathcal{P}_2}(\mathbf{x})\right| \leq e^d \cdot c_{46} \cdot \left(\frac{2a}{M^2}\right)^{p}.$$

A NN approximating $\widehat{T}_{f,q,\mathcal{P}_2}$ is also a good approximation for $T_{f,q,\mathcal{P}_2}(\mathbf{x})$.

Step 1*-2: A New Recursive Definition of $\widehat{T}_{f,q,\mathcal{P}}(\mathbf{x})$

Set

$$\phi_{1,0} = \mathbf{x}, \qquad \phi_{2,0} = \mathbf{0}_d, \qquad \phi_{3,0}^{(\mathbf{l})} = \phi_{4,0}^{(\mathbf{l})} = 0.$$

And for $j = 1, \ldots, M^d$, set

$$\phi_{1,j} = \phi_{1,j-1}, \qquad \phi_{2,j} = (C_{j,1})_{left} \cdot \mathbb{I}_{C_{j,1}}(\phi_{1,j-1}) + \phi_{2,j-1}$$

$$\phi_{3,j}^{(\mathbf{l})} = (\partial^{(\mathbf{l})} f)((C_{j,1})_{left}) \cdot \mathbb{I}_{C_{j,1}}(\phi_{1,j-1}) + \phi_{3,j-1}^{(\mathbf{l})}, \quad \phi_{4,j}^{(\mathbf{l})} = b_j^{(\mathbf{l})} \cdot \mathbb{I}_{C_{j,1}}(\phi_{1,j-1}) + \phi_{4,j-1}^{(\mathbf{l})}.$$

Furthermore, for $j = 1, \ldots, M^d$, set

$$\phi_{1,M^d+j} = \phi_{1,M^d+j-1}, \qquad \phi_{2,M^d+j} = \phi_{2,M^d+j-1} + \mathbf{v}_{j+1},$$

and

$$\phi_{3,M^d+j}^{(\mathbf{l})} = \sum_{\mathbf{s} \in \mathbb{N}^d : \|\mathbf{s}\|_1 \leq q - \|\mathbf{l}\|_1} \frac{\phi_{3,M^d+j-1}^{(\mathbf{l}+\mathbf{s})}}{\mathbf{s}!} \cdot (\mathbf{v}_{j+1})^{\mathbf{s}}$$

$$+ \left(\lfloor (4 + 2\lceil e^d \rceil) \cdot \phi_{4,M^d+j-1}^{(\mathbf{l})} \rfloor - \lceil e^d \rceil - 2 \right) \cdot c_{46} \cdot \left(\frac{2a}{M^2} \right)^{p - \|\mathbf{l}\|_1},$$

$$\phi_{4,M^d+j}^{(\mathbf{l})} = (4 + 2\lceil e^d \rceil) \cdot \phi_{4,M^d+j-1}^{(\mathbf{l})} - \lfloor (4 + 2\lceil e^d \rceil) \cdot \phi_{4,M^d+j-1}^{(\mathbf{l})} \rfloor.$$

Step 1*-2: A New Recursive Definition of $\widehat{T}_{f,q,\mathcal{P}}(\mathbf{x})$

And

$$\phi_{5,M^d+j} = \mathbb{I}_{\mathcal{A}^{(j)}}\left(\phi_{1,M^d+j-1}\right) \cdot \phi_{2,M^d+j-1} + \phi_{5,M^d+j-1}$$
$$\phi^{(\mathbf{l})}_{6,M^d+j} = \mathbb{I}_{\mathcal{A}^{(j)}}\left(\phi_{1,M^d+j-1}\right) \cdot \phi^{(\mathbf{l})}_{3,M^d+j-1} + \phi^{(\mathbf{l})}_{6,M^d+j-1}$$

where $\phi_{5,M^d j} = \mathbf{0}$, $\phi^{(\mathbf{l})}_{6,M^d} = 0$, and

$$\mathcal{A}^{(j)} = \Big\{ \mathbf{x} \in \mathbb{R}^d : -x_k + (\phi_{2,M^d+j-1})_k \leq 0,$$
$$\text{and } x_k - (\phi_{2,M^d+j-1})_k - \frac{2a}{M^2} < 0 \text{ for all } k = 1, \ldots, d \Big\}.$$

Then it can be shown that

$$\phi_{1,2M^d+1} = \sum_{\mathbf{l} \in \mathbb{N}^d : \|\mathbf{l}\|_1 \leq q} \frac{\phi^{\mathbf{l}}_{6,2M^d}}{\mathbf{l}!} \left(\phi_{1,2M^d} - \phi_{5,2M^d}\right)^{\mathbf{l}}$$

actually is $\widehat{T}_{f,q,\mathcal{P}_2}(\mathbf{x})$.

Step 2*: Approximating $\phi_{1,2M^d+1}$ by NN

Since the construction of $\phi_{1,2M^d+1}$ includes the function $\lfloor \cdot \rfloor$. We needs the following sub-lemma.

Lemma (sub-lemma 8)

Let $R > 0$, $B \in \mathbb{N}$, and

$$f_{ind,[j,\infty)}(z) = R \cdot \sigma(z-j) - R \cdot \sigma(z-j-1/R) \in \mathcal{F}(1,2)$$

with the ReLU activation function $\sigma(\cdot)$ for $j = 1, \ldots, B$. Then the NN

$$f_{trunc}(z) = \sum_{j=1}^{B} f_{ind,[j,\infty)}(z) \in \mathcal{F}(1, 2B)$$

satisfies $f_{trunc}(z) = \lfloor z \rfloor$ for $z \in [0, B+1)$ and $\min\{|z-j| : j \in \mathbb{N}\} \geq 1/R$.

With the assistance of sub-lemma 8, one can use the similar technique as we prove Approximation Lemma I to show that there exists a deep NN $f_{net,deep,\mathcal{P}_2}(\mathbf{x})$ with ReLU activation function can be enough close to $f(\mathbf{x})$ holds for all $\mathbf{x} \in \bigcup_{i=1}^{M^{2d}} (C_{i,2})^0_{1/M^{2p+2}}$, and totally bounded on $[-a, a)^d$.

Step B: Approximation of a hierarchical composition model

Recall the definition of hierarchical composition models, we denote $h_j^{(i)}$ as the j-th hierarchical composition model of some level i,

$$j = 1, \ldots, \widetilde{N}_i, \qquad i = 1, \ldots, l,$$

where $\widetilde{N}_i$ represents the minimal number of hierarchical composition models of level i. And it applies a $(p_j^{(i)}, C)$-smooth function $g_j^{(i)} : \mathbb{R}^{K_j^{(i)}} \to \mathbb{R}$ with $p_j^{(i)} = q_j^{(i)} + s_j^{(i)}$ with $(p_j^{(i)}, K_j^{(i)}) \in \mathcal{P}$. Then at the last level l, $h_1^{(l)}(\mathbf{x})$ can be recursively descrbied as follows:

$$h_j^{(i)}(\mathbf{x}) = g_j^{(i)} \left(h^{(i-1)}_{\sum_{t=1}^{j-1} K_t^{(i)}+1}(\mathbf{x}), \ldots, h^{(i-1)}_{\sum_{t=1}^{j} K_t^{(i)}}(\mathbf{x}) \right)$$

for $j = 1, \ldots, \widetilde{N}_i$ and $i = 2, \ldots, l$, and

$$h_j^{(1)}(\mathbf{x}) = g_j^{(1)} \left(x_{\pi(\sum_{t=1}^{j-1} K_t^{(1)}+1)}, \ldots, x_{\pi(\sum_{t=1}^{j-1} K_t^{(1)})} \right)$$

for some function $\pi : \{1, \ldots, \widetilde{N}_1\} \to \{1, \ldots, d\}$. Obviously, we have

$$\widetilde{N}_l = 1, \qquad \widetilde{N}_i = \sum_{j=1}^{\widetilde{N}_{i+1}} K_j^{(i+1)}, \qquad i = 1, \ldots, l-1.$$

An example of the structure $h_1^{(2)} \in \mathcal{H}(2, \mathcal{P})$

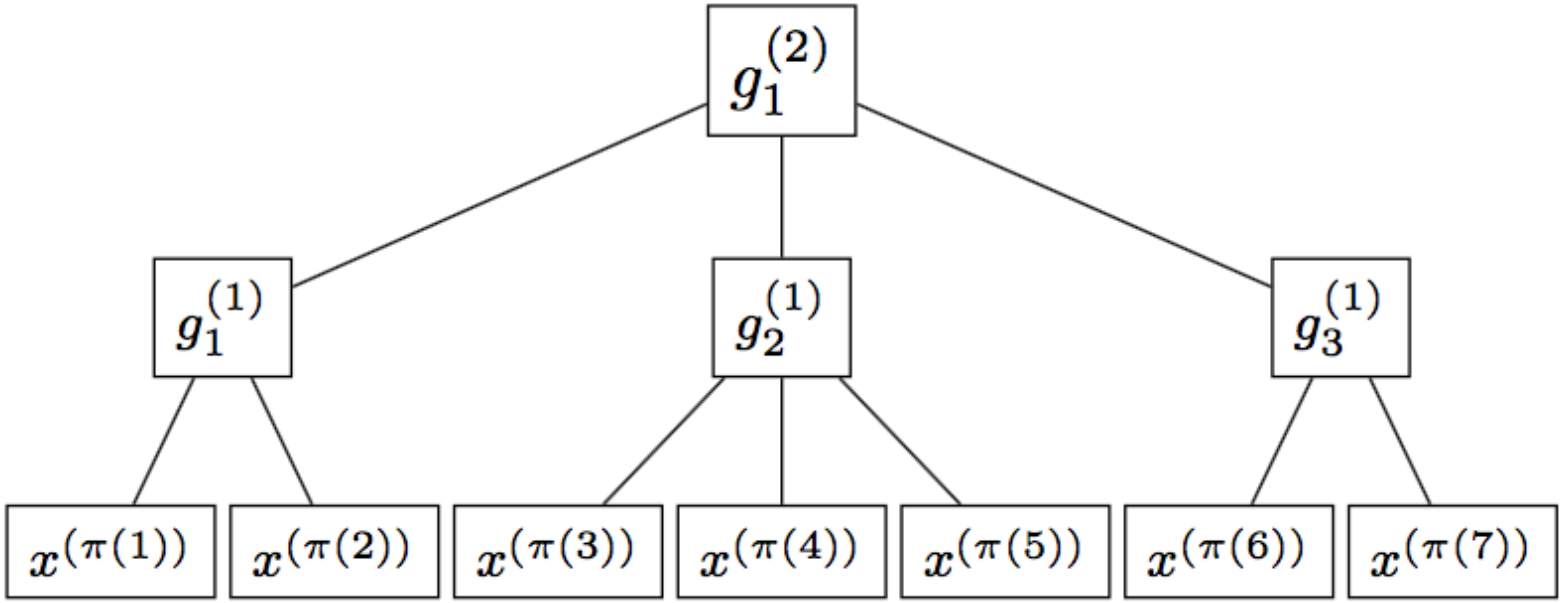


Illustration of a hierarchical composition model of the class $\mathcal{H}(2, \mathcal{P})$ with the structure $h_1^{(2)}(\mathbf{x}) = g_1^{(2)}(h_1^{(1)}(\mathbf{x}), h_2^{(1)}(\mathbf{x}), h_3^{(1)}(\mathbf{x}))$, $h_1^{(1)}(\mathbf{x}) = g_1^{(1)}(x_{\pi(1)}, x_{\pi(2)})$, $h_2^{(1)}(\mathbf{x}) = g_2^{(1)}(x_{\pi(3)}, x_{\pi(4)}, x_{\pi(5)})$, and $h_3^{(1)} = (x_{\pi(6)}, x_{\pi(7)})$

.

Step B: Approximation of a hierarchical composition model

Theorem (Approximation Theorem II)

Let $m : \mathbb{R}^d \to \mathbb{R}$ *be contained in the class* $\mathcal{H}(l, \mathcal{P})$ *for some* $l \in \mathbb{N}$ *and* $\mathcal{P} \subseteq [1, \infty) \times \mathbb{N}$. *Assume that the corresponding* $g_j^{(i)}$ *are Lipschitz continuous with Lipschitz constant* $C_{lip} \geq 1$ *and satisfies*

$$\|g_j^{(i)}\|_{C^{q_j^{(i)}}(\mathbb{R}^d)} \leq c_{20}$$

for some constant c_{20}. *Let* $a \geq 1$ *and* $M_{j,i} \in \mathbb{N}$ *independent with* a *sufficiently large with* $\min_{j,i} M_{j,i}^2 > c_{21} a^{4(p_{\max}+1)} / (2^l K_{\max} C_{Lip})^l$ *must hold for some* $c_{21} > 0$. *Then the in some NN class* $\mathcal{F}(L, r)$, *we can always find a good NN approximate* f *enough. Specially,Let* $L, r \in N$ *such that*

(a.i) $L \geq l\Big(5 + \lceil \log_4 \big(\max_{j,i} M_{j,i}^{2p_j^{(i)}} \big) \big(\log(K_{\max} \vee p_{\max} + 1) \rceil + 1 \big)\Big)$,

(a.ii) $r \geq \max_i \sum_{j=1}^{\widetilde{N}_j} 2^{K_j^{(i)}} \cdot 64 \cdot \binom{K_j^{(i)} + q_j^{(i)}}{K_j^{(i)}} \cdot (K_j^{(i)})^2 \cdot (q_j^{(i)} + 1) \cdot M_{j,i}^{K_j^{(i)}}$;

Step B: Approximation of a hierarchical composition model

Theorem (Approximation Theorem II, continue)

or

$$\text{(b.i)}\ L \geq \sum_{i=1}^{l} \sum_{j=1}^{\widetilde{N}_i} \Big(5M_{j,i}^{K_j^{(i)}} + \lceil \log_4 \big(M_{j,i}^{2p_j^{(i)}+4K_j^{(i)}(q_j^{(i)}+1)} \cdot e^{4(q_j^{(i)}+1)(M_{j,i}^{K_j^{(i)}}-1)} \big) \rceil \cdot \lceil \log_2(K_j^{(i)} \vee q_j^{(i)} + 1) + \lceil \log_4 M_{j,i}^{2p_j^{(i)}} \rceil,$$

$$\text{(b.ii)}\ r \geq 2\sum_{t=1}^{l-1} \widetilde{N}_t + 2d + 132 \cdot 2^{K_{\max}} \cdot \lceil e^{K_{\max}} \rceil \binom{K_{\max}+\lceil p_{\max} \rceil}{K_{\max}} \cdot \big((\lceil p_{\max} \rceil + 1) \vee K_{\max}^2 \big),$$

hold. There exists a NN $m_{NN} \in \mathcal{F}(L,r)$ *with property that*

$$\|m_{NN} - m\|_{\infty,[-a,a]^d} \lesssim a^{4(p_{\max}+1)} \max_{j,i} M_{j,i}^{-2p_j^{(i)}}.$$

The Proof of Approximation Theorem II

Recall the melting method (9), the basic idea is to define a composed network, which approximately computes $h_1^{(1)}, \ldots, h_{\widetilde{N}_1}^{(1)}, \ldots, h_1^{(l)}$.

Case a): Note that Approximation Theorem I a) guarantees there exist a wide NN can be close to $g_j^{(i)}$ enough, i.e.

$$f_{net,wide,g_j^{(i)}} \in \mathcal{F}(L_0, r_j^{(i)})$$

described in Approximation Theorem I a) with

$$L_0 = 5 + \lceil \log_4 \left(\max_{j,i} M_{j,i}^{2p_j^{(i)}} \right) \left(\log(K_{\max} \vee p_{\max} + 1) \rceil + 1 \right),$$

$$r_j^{(i)} = 2^{K_j^{(i)}} \cdot 64 \cdot \binom{K_j^{(i)} + q_j^{(i)}}{K_j^{(i)}} \cdot (K_j^{(i)})^2 \cdot (q_j^{(i)} + 1) \cdot M_{j,i}^{K_j^{(i)}}.$$

Under this case, it is easy to see the NN is contained in the class

$$\mathcal{F}\left(lL_0, \max_{i=1,\ldots,l} \sum_{j=1}^{\widetilde{N}_i} r_j^{(i)} \right) \subseteq \mathcal{F}(L, r).$$

The Proof of Approximation Theorem II

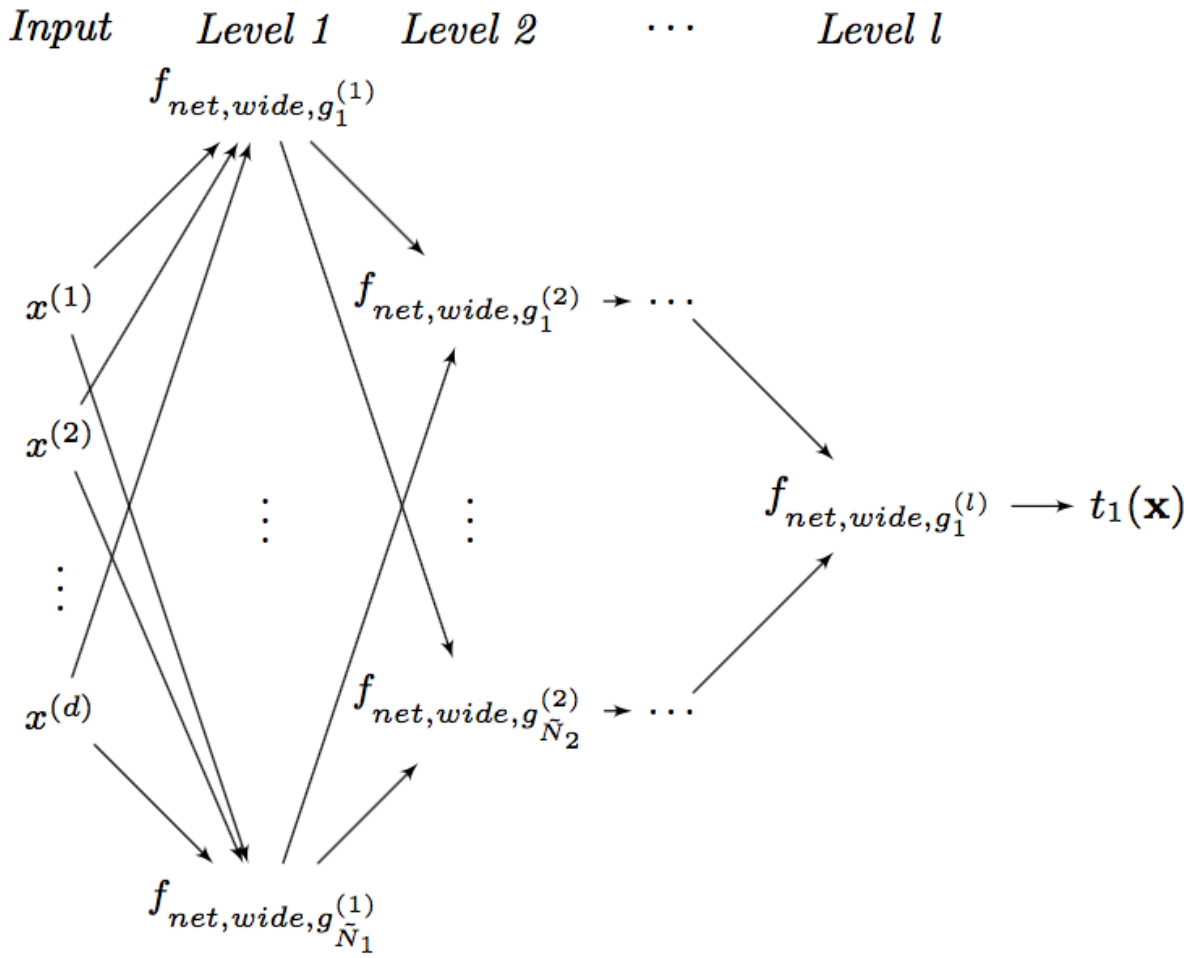


Illustration of the NN m_{NN} under case a).

The Proof of Approximation Theorem II

Case b): Deep NN require more layers. Denote $h_1^{(1)}, \ldots, h_{\tilde{N}_1}^{(1)}, \ldots, h_1^{(l)}$ by $h_1, \ldots, h_{\sum_{t=1}^{l} \tilde{N}_t}$.

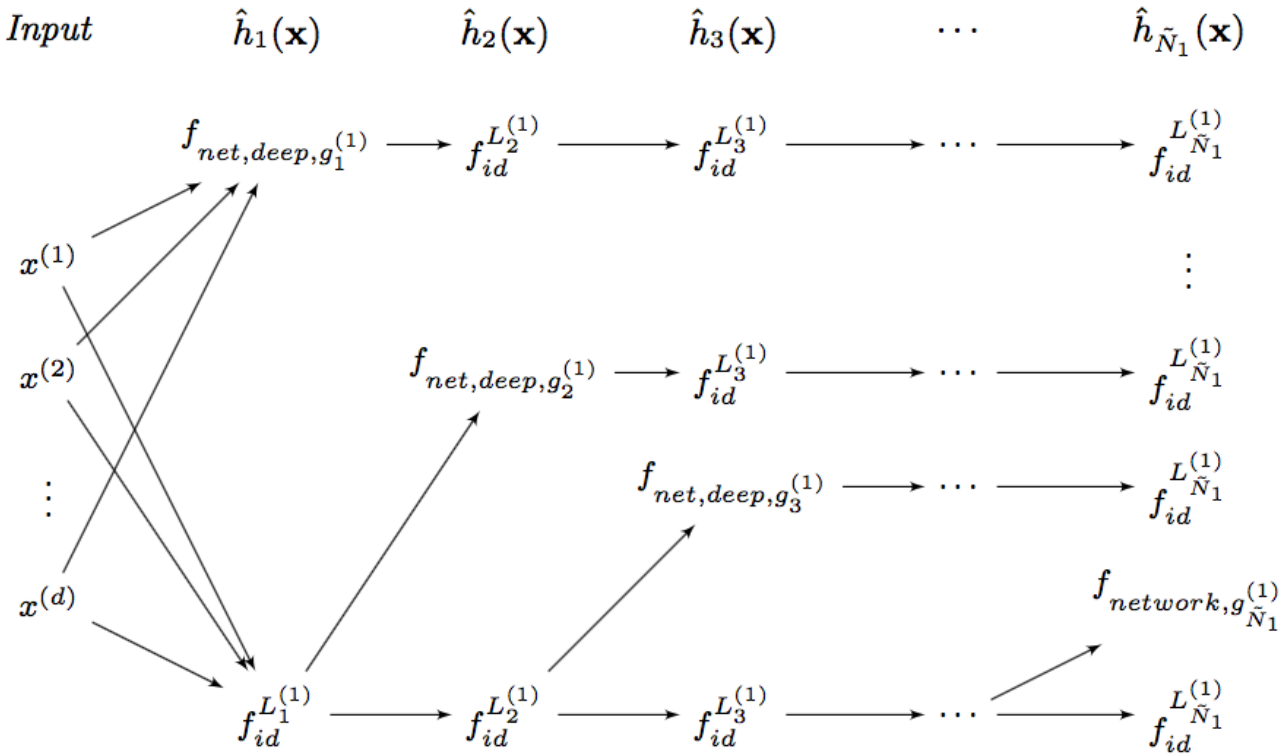


Illustration of the NN m_{NN} under case b).

The Proof of Approximation Theorem II

The $f_{net,deep,g_j^{(i)}}$ is in Approximation Theorem I b) satisfies

$$f_{net,deep,g_j^{(i)}} \in \mathcal{F}(L_j^{(i)}, r_0)$$

with

$$\begin{aligned} L_j^{(i)} =& 5M_{j,i}^{K_j^{(i)}} + \lceil \log_4 \left(M_{j,i}^{2p_j^{(i)}+4K_j^{(i)}(q_j^{(i)}+1)} \cdot e^{4(q_j^{(i)}+1)(M_{j,i}^{K_j^{(i)}}-1)} \right) \rceil \\ & \cdot \lceil \log_2(K_j^{(i)} \vee q_j^{(i)} + 1) + \lceil \log_4 M_{j,i}^{2p_j^{(i)}} \rceil, \\ r_0 =& 132 \cdot 2^{K_{\max}} \cdot \lceil e^{K_{\max}} \rceil \binom{K_{\max}+\lceil p_{\max} \rceil}{K_{\max}} \cdot \left((\lceil p_{\max} \rceil + 1) \vee K_{\max}^2 \right). \end{aligned}$$

Note that we also need $2d$ neurons per layer to successively apply f_{id} to the input $\mathbf{x}$ and 2 neurons per layer to apply f_{id} to the at most $\sum_{t=1}^{l-1} \widetilde{N}_t$ already computed functions in our NN. Hence m_{NN} is contained in the class

$$\mathcal{F}\left(\sum_{j=1}^{\widetilde{N}_j} \widetilde{L}_j, \, 2 \sum_{t=1}^{l-1} \widetilde{N}_t + 2d + r_0 \right),$$

with $\widetilde{L}_{N_j^{(i)}} = L_j^{(i)}$.

Step C: Estimator (2) is a good measure for the NN Step B.

Lemma (expected L_2-error of the LSE)

Assume that the distribution of (X, Y) satisfies $\mathrm{E}\exp\{c_1 Y^2\} < \infty$ for some constant $c_1 > 0$ and the regression function m is bounded in absolute value. Let $\widetilde{m}_n$ be the least squares estimator in (2) based on some function space $\mathcal{F}_n$ and set $m_n = T_{c_2 \cdot \log n} \widetilde{m}_n$ for some constant c_2. Then m_n satisfies

$$\begin{aligned}
&\mathrm{E} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \, P_X(d\mathbf{x}) \\
&\leq c_3 \frac{\log^2 n}{n} \sup_{\mathbf{x}_1^n \in (\mathbb{R}^d)^n} \log \left(\mathcal{N}\left(\frac{1}{c_2 n \log n}, T_{c_2 \log n} \mathcal{F}_n(\mathbf{x}_1^n), \mathbb{P}_n \right) + 1 \right) \\
&\quad + 2 \inf_{f \in \mathcal{F}_n} \int |f(\mathbf{x}) - m(\mathbf{x})|^2 \, P_X(d\mathbf{x})
\end{aligned}$$

for $n > 1$ and some constant $c_3 > 0$, which does not depend on n or the parameters in the estimate.

The proof can be seen in [8].

Step C: Estimator (2) is a good measure for the NN Step B.

The second term in the left side of the above inequality can be bounded by

$$\asymp \left(a_n^{4(p_{\max}+1)}\right)^2 \cdot \max_{j,i} M_{j,i}^{-4p_j^{(i)}} \asymp \log^2 n \max_{j,i} n^{-\frac{2p_j^{(i)}}{2p_j^{(i)}+K_j^{(i)}}}$$

by choosing

$$M_{j,i} = \left\lceil n^{\frac{1}{2(2p_j^{(i)})+K_j^{(i)}}} \right\rceil, \qquad a_n = (\log n)^{\frac{1}{4(p_{\max}+1)}}.$$

And the covering number in the first term in the left side of the above inequality can also be bounded by the following lemma.

Lemma (a bound on the covering number)

Let $1/n^{c_1} \leq \epsilon \leq c_2 \log n/8$ *with certain constants* $c_1, c_2 > 0$. *Let* $L, r \in \mathbb{N}$, *then*

$$\log \mathcal{N}\left(\epsilon,\, T_{c_2 \log n}\mathcal{F}(L,r)(\mathbf{x}_1^n),\, \mathbb{P}_n\right) \leq c_3 \log n \cdot L^2 r^2 \log(Lr^2)$$

holds for sufficiently large n, $\mathbf{x}_1^n \in (\mathbb{R}^d)^n$, *and the constant* $c_3 > 0$ *is independent of* n, L, *and* r.

Reference I

[1] Charles J Stone. "Optimal global rates of convergence for nonparametric regression". In: *The annals of statistics* (1982), pp. 1040–1053.

[2] László Györfi et al. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[3] Joel L Horowitz, Enno Mammen, et al. "Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions". In: *The Annals of Statistics* 35.6 (2007), pp. 2589–2619.

[4] Andrew R Barron. "Approximation and estimation bounds for artificial neural networks". In: *Machine learning* 14.1 (1994), pp. 115–133.

[5] Daniel F McCaffrey and A Ronald Gallant. "Convergence rates for single hidden layer feedforward networks". In: *Neural Networks* 7.1 (1994), pp. 147–158.

Reference II

[6] Michael Kohler and Adam Krzyżak. "Nonparametric regression based on hierarchical interaction models". In: *IEEE Transactions on Information Theory* 63.3 (2016), pp. 1620–1630.

[7] Michael Kohler and Adam Krzyżak. "Adaptive regression estimation with multilayer feedforward neural networks". In: *Nonparametric Statistics* 17.8 (2005), pp. 891–913.

[8] Benedikt Bauer, Michael Kohler, et al. "On deep learning as a remedy for the curse of dimensionality in nonparametric regression". In: *Annals of Statistics* 47.4 (2019), pp. 2261–2285.

[9] Johannes Schmidt-Hieber et al. "Nonparametric regression using deep neural networks with ReLU activation function". In: *Annals of Statistics* 48.4 (2020), pp. 1875–1897.

Thank You!